NYU

DS-GA 3001.009 Responsible Data Science Lab 12

Center for Data Science
Tandon School of Engineering

Outline

Interpretability

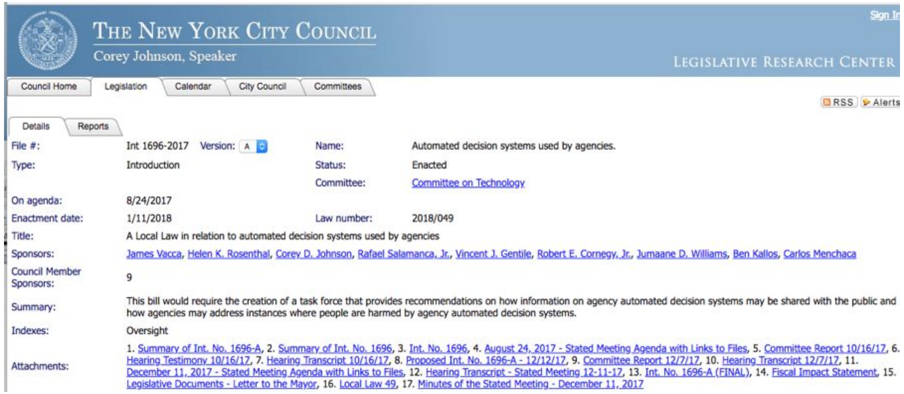
Nutritional Labels for rankings

Other frameworks

- **Data Sheets**
- **Model Cards**

Project

Interpretability



Explain assumptions and effects

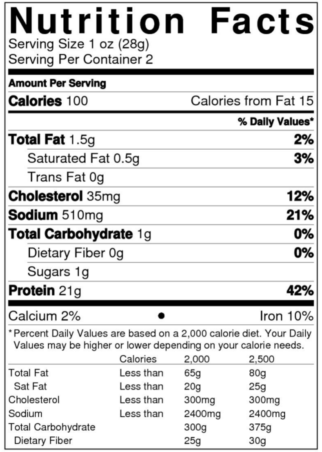
- not details of operation
- help users understand

Engage the public

- technical and non-technical

Interpretability *at every stage of the data lifecycle*

- useful internally during development
- communication and coordination between agencies
- accountability to the public

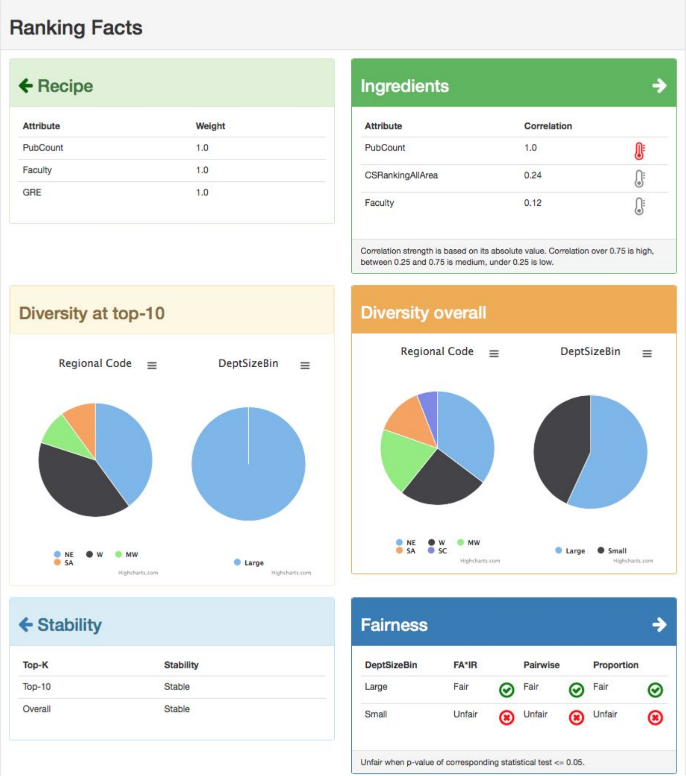


Nutrition Labels for rankings

Yang, K., Stoyanovich, J., Asudeh, A., Howe, B., Jagadish, H. V., & Miklau, G. (2018, May). A nutritional label for rankings. In Proceedings of the 2018 International Conference on Management of Data (pp. 1773-1776). ACM.

An interpretability tool “*Ranking Facts*”
- simple and standardized labels

Explains **ranked outputs** to users
- Provides summarized information regarding ranking process
- Includes interpretations of fairness, stability, and transparency for ranked outputs



Nutrition Labels for rankings

Web-based application

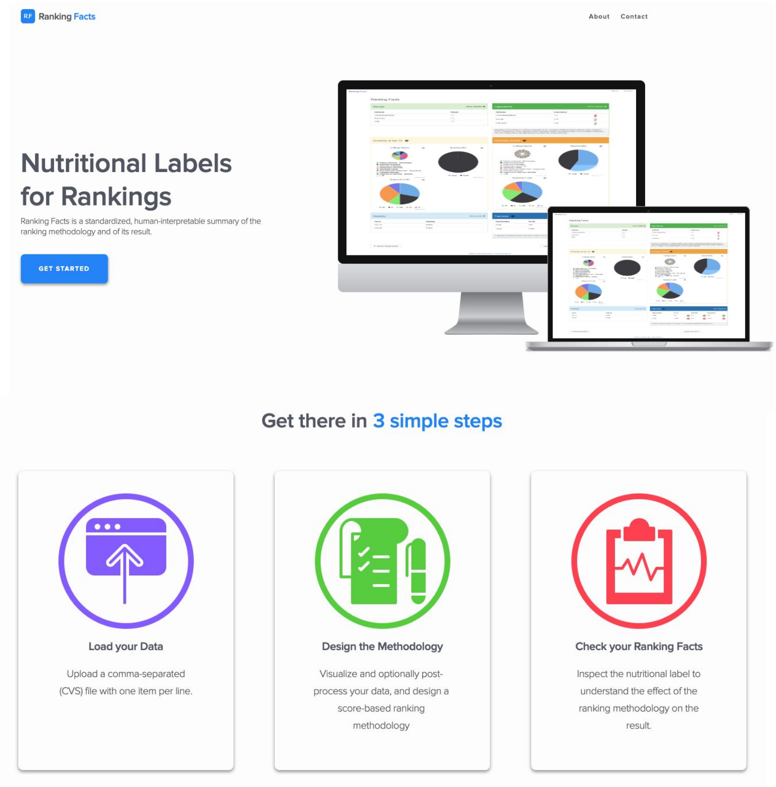
http://demo.dataresponsibly.com/rankingfacts/

Support users' own datasets

Automated label generation - unlike other methods we'll discuss today

Focus on algorithmic **ranker**

- A rule-based system, not machine learning
- e.g., college rankings
  - ranking methodology is inspired by US World & News Report and CS rankings



Data Sheets

Gebru, Timnit, et al. "Datasheets for datasets." arXiv preprint arXiv:1803.09010 (2018).

- Motivation for dataset creation
- Composition of the dataset
- Data collection process
- Pre-processing of the data
- Distribution of the data
- Maintenance of the data
- Legal and ethical considerations

Legal & Ethical Considerations

If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection? (e.g., datasets that collect writing, photos, interactions, transactions, etc.)

If it relates to other ethically protected subjects, have appropriate obligations been met? (e.g., medical data might include information collected from animals)

If it relates to people, were there any ethical review applications/reviews/approvals? (e.g. Institutional Review Board applications)

If it relates to people, were they told what the dataset would be used for and did they consent? What community norms exist for data collected from human communications? If consent was obtained, how? Were the people provided with any mechanism to revoke their consent in the future or for certain uses?

If it relates to people, could this dataset expose people to harm or legal action? (e.g., financial social or otherwise) What was done to mitigate or reduce the potential for harm?

If it relates to people, does it unfairly advantage or disadvantage a particular social group? In what ways? How was this mitigated?

If it relates to people, were they provided with privacy guarantees? If so, what guarantees and how are these ensured?

Does the dataset comply with the EU General Data Protection Regulation (GDPR)? Does it comply with any other standards, such as the US Equal Employment Opportunity Act?

Motivation for Dataset Creation

Why was the dataset created? (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)

What (other) tasks could the dataset be used for? Are there obvious tasks for which it should *not* be used?

Has the dataset been used for any tasks already? If so, where are the results so others can compare (e.g., links to published papers)?

Who funded the creation of the dataset? If there is an associated grant, provide the grant number.

Any other comments?

Dataset Distribution

How is the dataset distributed? (e.g., website, API, etc.; does the data have a DOI; is it archived redundantly?)

When will the dataset be released/first distributed? (Is there a canonical paper/reference for this dataset?)

What license (if any) is it distributed under? Are there any copyrights on the data?

Are there any fees or access/export restrictions?

Any other comments?

Data Sheets

Dataset: http://vis-www.cs.umass.edu/lfw/

Motivation for Dataset Creation

Why was the dataset created? (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)

Labeled Faces in the Wild was created to provide images that can be used to study face recognition in the unconstrained setting where image characteristics (such as pose, illumination, resolution, focus), subject demographic makeup (such as age, gender, race) or appearance (such as hairstyle, makeup, clothing) cannot be controlled. The dataset was created for the specific task of pair matching: given a pair of images each containing a face, determine whether or not the images are of the same person.[1]

What (other) tasks could the dataset be used for? Are there obvious tasks for which it should *not* be used?

The LFW dataset can be used for the face identification problem. Some researchers have developed protocols to use the images in the LFW dataset for face identification.[2]

Has the dataset been used for any tasks already? If so, where are the results so others can compare (e.g., links to published papers)?

Papers using this dataset and the specified evaluation protocol are listed in http://vis-www.cs.umass.edu/lfw/results.html

Who funded the creation of the dataset? If there is an associated grant, provide the grant number.

The building of the LFW database was supported by a United States National Science Foundation CAREER Award.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Each instance is a pair of images labeled with the name of the person in the image. Some images contain more than one face. The labeled face is the one containing the central pixel of the image—other faces should be ignored as "background".

How many instances are there in total (of each type, if appropriate)?

The dataset consists of 13,233 face images in total of 5749 unique individuals. 1680 of these subjects have two or more images and 4069 have single ones.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

Data Sheets

Dataset: http://vis-www.cs.umass.edu/lfw/

Data Preprocessing

What preprocessing/cleaning was done? (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values, etc.)

The following steps were taken to process the data:

1. **Gathering raw images:** First the raw images for this dataset were obtained from the Faces in the Wild dataset consisting of images and associated captions gathered from news articles found on the web.

2. **Running the Viola-Jones face detector**[5] The OpenCV version 1.0.0 release 1 implementation of Viola-Jones face detector was used to detect faces in each of these images, using the function cvHaarDetectObjects, with the provided Haar classifier—cascadehaarcascadefrontalfacedefault.xml. The scale factor was set to 1.2, min neighbors was set to 2, and the flag was set to CV HAAR DO CANNY PRUNING.

3. **Manually eliminating false positives:** If a face was detected and the specified region was determined not to be a

Dataset Distribution

How is the dataset distributed? (e.g., website, API, etc.; does the data have a DOI; is it archived redundantly?)

The dataset can be downloaded from http://vis-www.cs.umass.edu/lfw/index.html#download. The images can be downloaded as a gzipped tar file.

When will the dataset be released/first distributed? (Is there a canonical paper/reference for this dataset?)

The dataset was released in October, 2007.

What license (if any) is it distributed under? Are there any copyrights on the data?

The crawled data copyright belongs to the news papers that the data originally appeared in. There is no license, but there is a request to cite the corresponding paper if the dataset is used: Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. University of Massachusetts, Amherst, Technical Report 07-49, October, 2007.

What license (if any) is it distributed under? Are there any copyrights on the data?

Are there any fees or access/export restrictions?

There are no fees or restrictions.

Dataset Nutrition Label

Web version: https://ahmedhosny.github.io/datanutrition/

Dataset Facts

ProPublica's Dollars for Docs Data

Metadata

Filename	201612v1-docdollars-product_payments
Format	csv
Url	https://projects.propublica.org/docdollars/
Domain	healthcare
Keywords	Physicians, drugs, medicine, pharmaceutical, transactions
Type	tabular
Rows	500
Columns	18
Missing	5.2%
License	cc
Released	JAN 2017
Range	
From	AUG 2013
To	DEC 2015

Description This is the data used in ProPublica's Dollars for Docs news application. It is primarily based on CMS's Open Payments data, but we have added a few features. ProPublica has standardized drug, device and manufacturer names, and made a flattened table (product_payments) that allows for easier aggregating payments associated with each drug/device. In [1], one payment record can be attributed to up to five different drugs or medical devices. This table flattens the payments out so that each drug/device related to each payment gets its own line.

Provenance

Source

Name	U.S. Centers for Medicare & Medicaid Services
Url	https://www.cms.gov/OpenPayments/
Email	openpayments@cms.hhs.gov

Author

Name	Propublica
Url	https://www.propublica.org/datastore/
Email	data.store@propublica.org

Statistics

Ordinal

name	type	count	uniqueEntries	mostFrequent	leastFrequent	missing
id	number	500	488 including missing	missing value (13)	multiple detected	2.60%
applicable_manufacturer_or_app…	number	500	4	100000000232 (417)	multiple detected	0%
date_of_payment	date	500	213 including missing	missing value (27)	multiple detected	5.40%
general_transaction_id	number	500	467 including missing	missing value (34)	multiple detected	6.80%
program_year	number	500	2 including missing	2014 (495)	missing value (5)	1.00%

Nominal

name	type	count	uniqueEntries	mostFrequent	leastFrequent	missing
product_name	string	500	16 including missing	Xarelto (200)	Aciphex (1)	3.20%
original_product_name	string	500	15	Xarelto (212)	Aciphex (1)	0%
product_ndc	number	500	21 including missing	5045857810 (201)	multiple detected	5.00%
product_is_drug	boolean	500	2 including missing	t (492)	missing value (8)	1.60%
payment_has_many	boolean	500	3 including missing	f (267)	missing value (29)	5.80%
teaching_hospital_id	number	500	2 including missing	0 (464)	missing value (36)	7.20%
physician_profile_id	number	500	230 including missing	missing value (32)	multiple detected	6.40%
recipient_state	string	500	40	CA (56)	multiple detected	0%
applicable_manufacturer_or_app…	string	500	5 including missing	Janssen Pharmaceuticals, Inc (3…	multiple detected	7.00%
teaching_hospital_ccn	number	500	2 including missing	0 (481)	missing value (19)	3.80%
product_slug	string	500	15 including missing	drug-xarelto (196)	drug-aciphex (1)	8.20%

Continuous

name	type	count	min	median	max	mean	standardDeviation	missing	zeros
total_amount_of_pay…	number	500	0.14	14.00	5000	134.21	501.99	9.40%	0%

Discrete

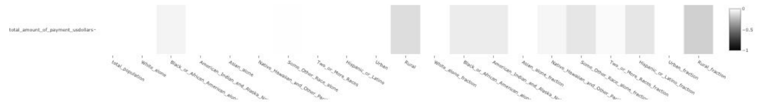
name	type	count	min	median	max	mean	standardDeviation	missing	zeros
number_of_payments…	number	500	1	1.00	1	1.00	0.00	4.80%	0%

Ground Truth Correlations

negative correlation

Variable

total_amount_of_payment_usdollars



Model Cards

Mitchell, Margaret, et al. "Model cards for model reporting." Proceedings of the Conference on Fairness, Accountability, and Transparency. ACM, 2019.

- ML and AI practitioners
- Model developers
- Software developers
- Policymakers
- Organizations
- ML-knowledgeable individuals

Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Model Cards

Model Card - Toxicity in Text

Model Details

- The TOXICITY classifier provided by Perspective API [32], trained to predict the likelihood that a comment will be perceived as toxic.
- Convolutional Neural Network.
- Developed by Jigsaw in 2017.

Intended Use

- Intended to be used for a wide range of use cases such as supporting human moderation and providing feedback to comment authors.
- Not intended for fully automated moderation.
- Not intended to make judgments about specific individuals.

Factors

- Identity terms referencing frequently attacked groups, focusing on sexual orientation, gender identity, and race.

Metrics

- Pinned AUC, as presented in [11], which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups.

Ethical Considerations

- Following [31], the Perspective API uses a set of values to guide their work. These values are Community, Transparency, Inclusivity, Privacy, and Topic-neutrality. Because of privacy considerations, the model does not take into account user history when making judgments about toxicity.

Training Data

- Proprietary from Perspective API. Following details in [11] and [32], this includes comments from a online forums such as Wikipedia and New York Times, with crowdsourced labels of whether the comment is "toxic".
- "Toxic" is defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."

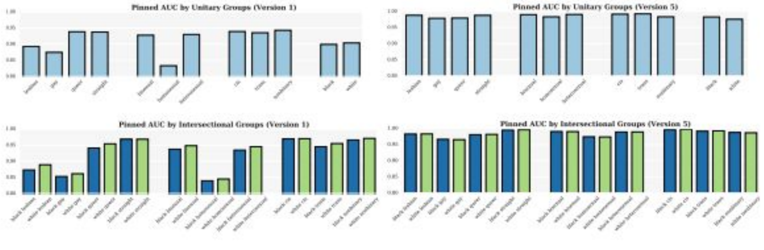
Evaluation Data

- A synthetic test set generated using a template-based approach, as suggested in [11], where identity terms are swapped into a variety of template sentences.
- Synthetic data is valuable here because [11] shows that real data often has disproportionate amounts of toxicity directed at specific groups. Synthetic data ensures that we evaluate on data that represents both toxic and non-toxic statements referencing a variety of groups.

Caveats and Recommendations

- Synthetic test data covers only a small set of very specific comments. While these are designed to be representative of common use cases and concerns, it is not comprehensive.

Quantitative Analyses



Model Card - Smiling Detection in Images

Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors

- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics

- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Training Data

- CelebA [36], training data split.

Evaluation Data

- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

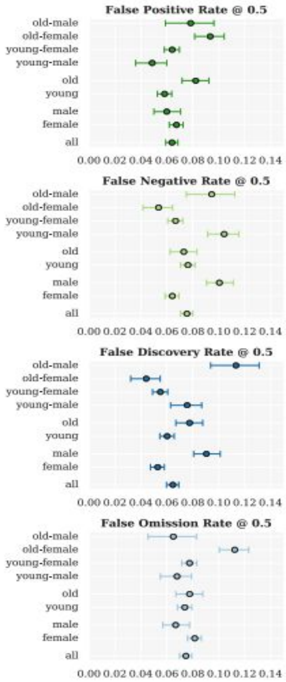
Ethical Considerations

- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Caveats and Recommendations

- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Quantitative Analyses



Project - Nutrition Labels for ADS

Data

- Data profiling
- Data preprocessing
- ...

→ *NYC open data portal*

→ *complete Kaggle competition*

ADS

- Black-box systems?
- How are they used?
- ...

→ *AI NOW NYC ADS charts*

https://ainowinstitute.org/nycadschart.pdf

→ *completed Kaggle competitions*

https://www.kaggle.com/competitions

KNOWN NEW YORK CITY USE CASES

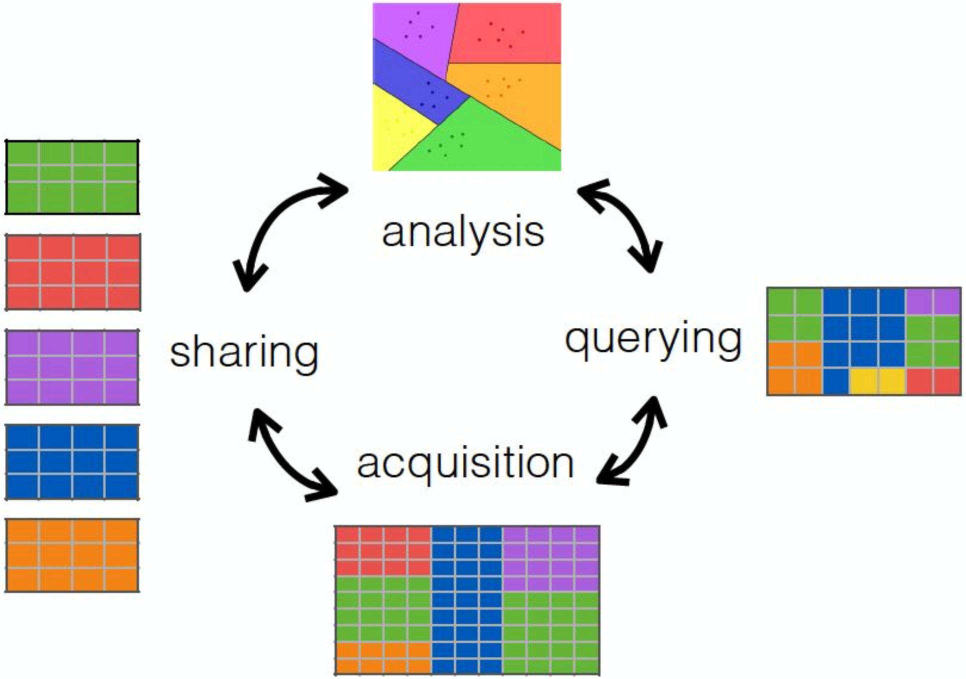
Issue	Description of Decision System	Links for Examples
Child Welfare	**Child Risk and Safety Assessments** are used by child welfare agencies to evaluate potential child neglect and abuse cases for risk of child death/injury. Data often comes from multiple sources, including a jurisdiction's department of human services and the police. They are often not designed to give ultimate decisions on child placement, but to advise on whether a reported case of potential child abuse/neglect should be further investigated or reviewed.	Chicago failed example Allegheny County example NYC example
Criminal Justice	**DNA Analysis**, also known as probabilistic genotyping, these systems interpret forensic DNA samples by performing statistical analysis on a mixture of DNA from different people to determine the probability that a sample is from a potential suspect.	TrueAllele NYC example
	Inmate Housing Classification is a system that analyzes a variety of criminal justice data and outcomes to determine the conditions of confinement, eligibility for programming, and overall housing arrangements of inmates in a jail or prison.	NYC example California study Study of Pennsylvania system

AI NOW's NYC ADC charts

Project - Nutrition Labels for ADS

Outcomes

- Fairness
- Diversity
- Transparency
- ...



Project

Proposal (due **5pm, Apr 29**)

- data and ADS

Notebook (due **11 am, May 13**)

- includes all interpretability components

Report (due **11 am, May 13**)

Presentation (11 am, May 13)

- 5 mins

Thank you!